Computing Quality Scores and Uncertainty for Approximate Pattern Matching in Geospatial Semantic Graphs

Randy C. Brost, Cynthia A. Phillips, David G. Robinson, David J. Stracuzzi, and Diane Myung-kyung Suh

Summary

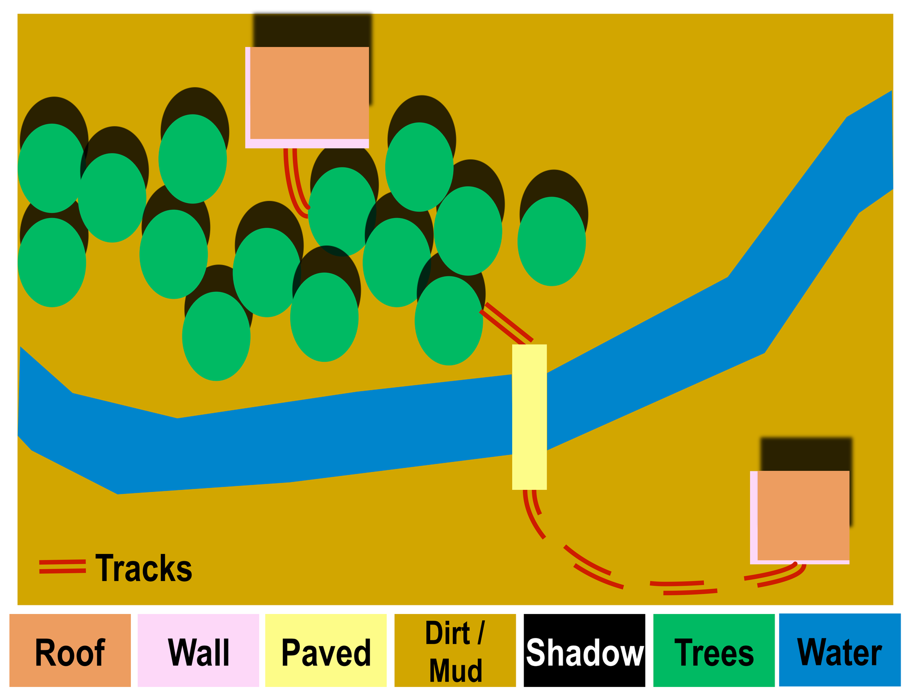
Geospatial semantic graphs provide a robust foundation for representing and analyzing remote sensor data. In particular, semantic graphs support a variety of pattern search operations that capture the spatial and temporal relationships among the objects and events in the data. However, in the presence of large data corpora, even a carefully constructed search query can return a large number of unanticipated or spurious matches. This work considers the problem of calculating a quality score for each match to the query, given that the underlying data are uncertain. We present preliminary algorithms for determining both match quality scores and associated uncertainty bounds, illustrated in the context of an example problem.

Finding Activities of Interest in Imagery

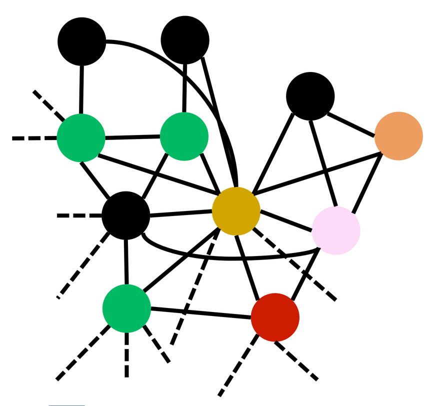
1. Collect data

Feature Extraction
Image Segmentation
Segment Classification

2. Compute semantic interpretation

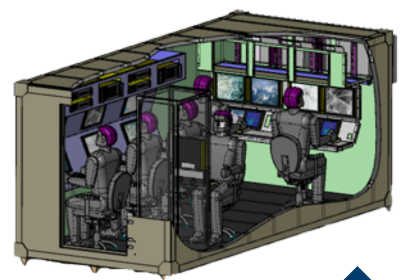


3. Construct a semantic graph

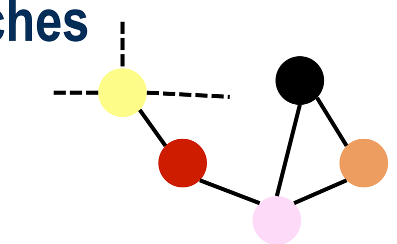


4. Query the graph

Find little red riding hood's drive to grandma's house.



5. Render candidate matches



Given: Candidate matches to a query pattern. **Produce:** Match probability/quality scores with associated uncertainty bounds.

Two-Stage Approach

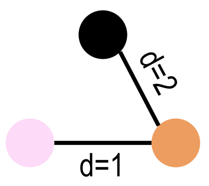
We assume very primitive semantic classes that maximize the reliability of the initial interpretation of the sensor data. Different sensors have unique strengths, and therefore produce different primitive semantics. Our goal is then to overlay a semantic hierarchy onto the semantic graph to better support search and analysis.

1. For commonly-occurring classes, use supervised data to compute the probability that candidate objects match a complex semantic pattern.
2. For rare classes, compute similarity of candidates to the prescriptive query.

This combined approach lets us use domain expertise and background knowledge where available, while remaining flexible to situations in which it is not.

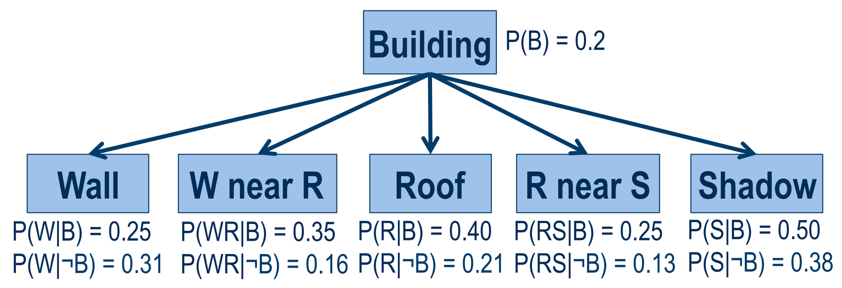
Hierarchical Naïve Bayesian Classifier

Goal: Is this a building?



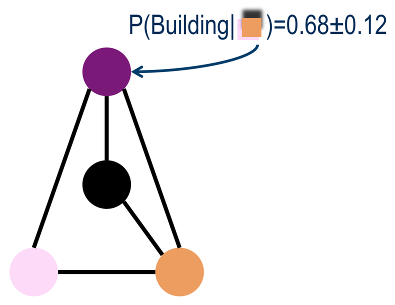
1. Subgraph associated with primitive semantic features

2. Naïve Bayesian classifier with confidence interval estimate

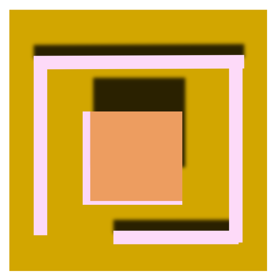


$$P(c=\text{Building} \mid X=\blacksquare) = \frac{P(c)\prod_{i=1}^{|A|} P(a_i \mid c)}{P(X)}$$

$$CI\left[P(c)\prod_{i=1}^{|A|} P(a_i \mid c)\right] = P(c \mid X) \pm z_{1-\frac{\alpha}{2}} \operatorname{std}\left(P(c)\prod_{i=1}^{|A|} P(a_i \mid c)\right)$$

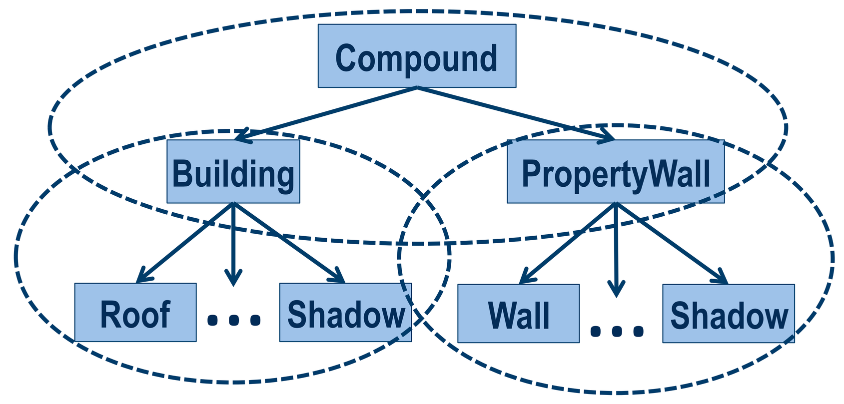


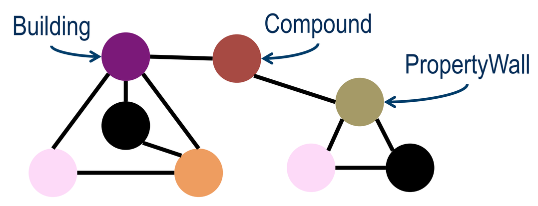
3. Add "building" to subgraph



4. Now consider "compound"

5. Estimate each object independently





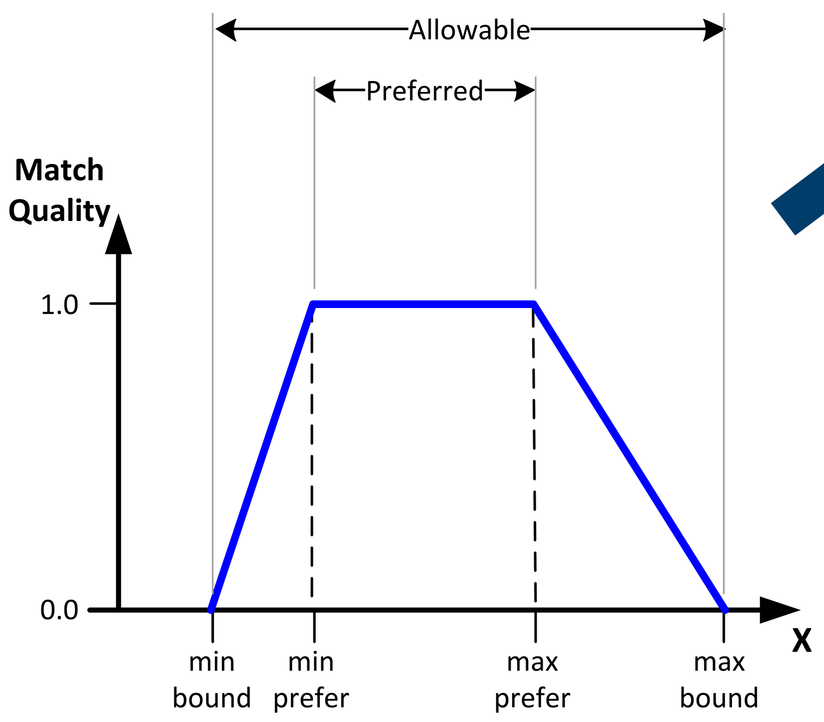
6. Resulting subgraph

Similarity Measures

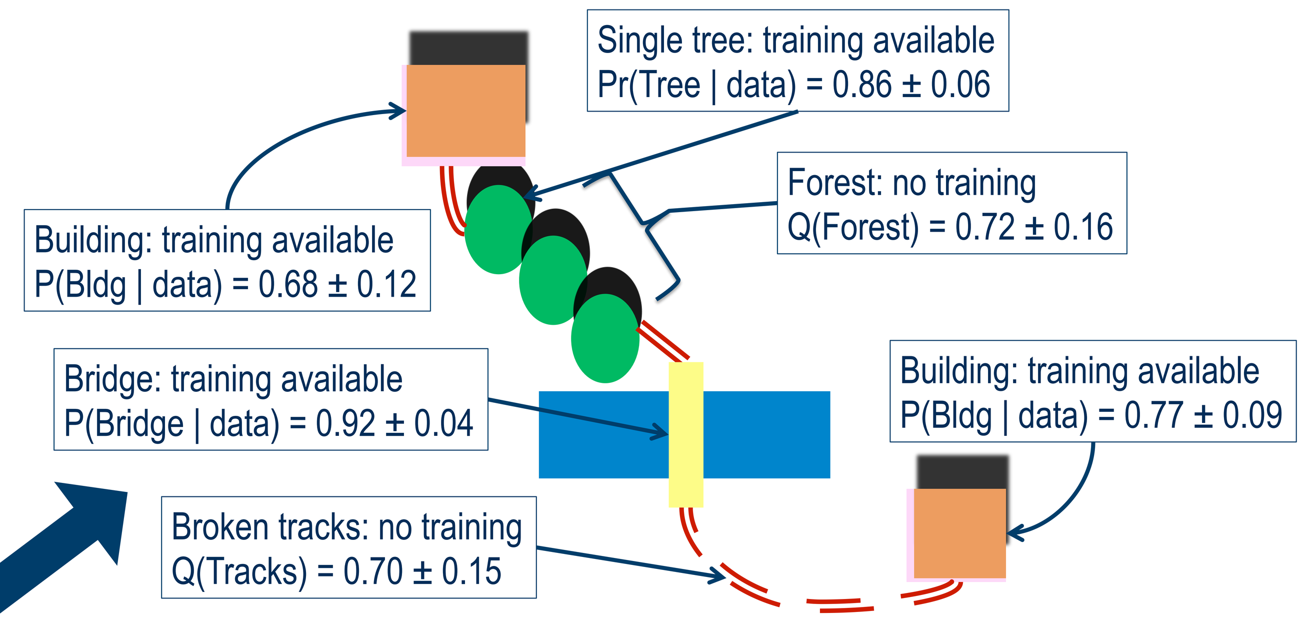
Goal: Find little red riding hood's drive to grandma's house.

- *Vehicle tracks lead from a building*
- *Over a bridge*
- *Through a forest*
- *To a second building*

1. Scores based on match of node attributes to specified ideals

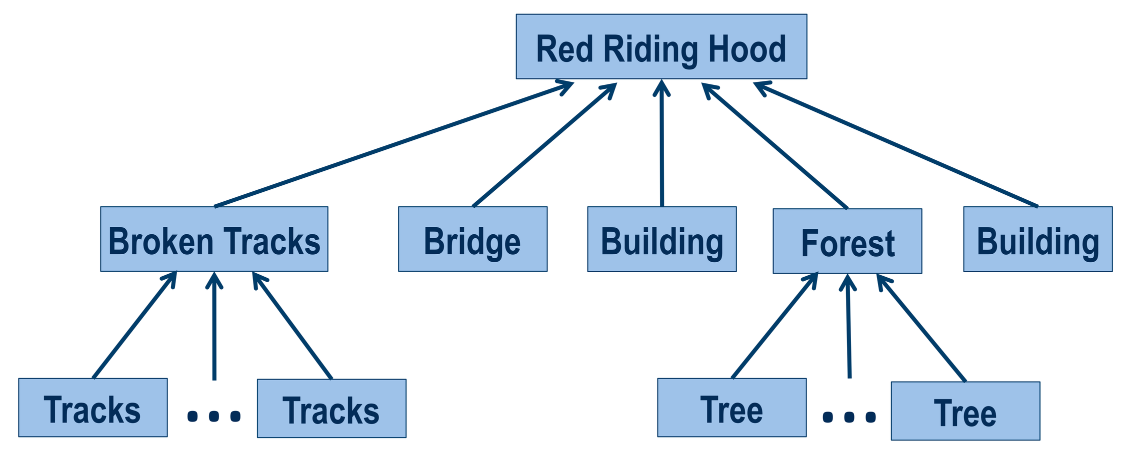


2. Evaluate quality of each component
   - Use Naïve Bayes when data is available
   - Otherwise similarity measures



3. Aggregate quality over node components

$$Q(\text{node}) = \sum_i k_i \sqrt{\prod_{a_i \in \text{node}} Q(a_i)^{k_i}}$$



4. Bootstrap confidence intervals

Next Steps

- Test described methods with a large corpus of automatically labeled sensor data (nearly complete).
- Improve confidence interval estimates for naïve Bayes. Currently, they account for randomness in the data, but not for randomness in the model parameters. As a result, the confidence intervals are too narrow, and may be biased.
- Consider variations of the naïve Bayesian calculation that propagate probabilistic information from low-level components to higher-level components.
- Investigate other similarity measures. The current measure is based on the geometric mean, but other measures, such as the generalized mean, may be more appropriate.
- Continue to investigate other approaches to computing match quality scores and confidence intervals. A variety of methods are available, each with a unique trade-off between required background knowledge and theoretical justification.

Key References

- Langley, P. (1993). Induction of recursive Bayesian classifiers. In Machine Learning: ECML-93 (pp. 153-164). Springer, Berlin Heidelberg.
- Pronk, V., Gutta, S.V.R., Verhaegh, W.F.J. (2005). Incorporating confidence in a naïve Bayes classifier. User Modeling 2005, LNAI 3538, (pp. 317–326).

Acknowledgements

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

SAND Number: 2014-1753P
Presented at: Conference on Data Analysis (CoDA) 2014

Contact: David J. Stracuzzi <djstrac@sandia.gov>